An affine invariant interest point detector

Krystian Mikolajczyk and Cordelia Schmid

INRIA Rhône-Alpes & GRAVIR-CNRS
655, av. de l'Europe, 38330 Montbonnot, France
{Name.Surname}@inrialpes.fr
http://www.inrialpes.fr/movi

Abstract. This paper presents a novel approach for detecting affine invariant interest points. Our method can deal with significant affine transformations including large scale changes. Such transformations introduce significant changes in the point location as well as in the scale and the shape of the neighbourhood of an interest point. Our approach allows to solve for these problems simultaneously. It is based on three key ideas: 1) The second moment matrix computed in a point can be used to normalize a region in an affine invariant way (skew and stretch). 2) The scale of the local structure is indicated by local extrema of normalized derivatives over scale. 3) An affine-adapted Harris detector determines the location of interest points. A multi-scale version of this detector is used for initialization. An iterative algorithm then modifies location, scale and neighbourhood of each point and converges to affine invariant points. For matching and recognition, the image is characterized by a set of affine invariant points; the affine transformation associated with each point allows the computation of an affine invariant descriptor which is also invariant to affine illumination changes. A quantitative comparison of our detector with existing ones shows a significant improvement in the presence of large affine deformations. Experimental results for wide baseline matching show an excellent performance in the presence of large perspective transformations including significant scale changes. Results for recognition are very good for a database with more than 5000 images.

Keywords : Image features, matching, recognition.

1 Introduction

Local characteristics have shown to be well adapted to matching and recognition, as they allow robustness to partial visibility and clutter. The difficulty is to obtain invariance under arbitrary viewing conditions. Different solutions to this problem have been developed over the past few years and are reviewed in section 1.1. These approaches first detect features and then compute a set of descriptors for these features. They either extract invariant features (and descriptors) or they compute invariant descriptors based on non-invariant features. In the case of significant transformations feature detection has to be adapted to the transformation, as at least a subset of the features must be present in both images in order to allow for correspondences. Features which have shown to be particularly appropriate are interest points. Scale invariant interest points detectors have been presented previously [10, 11]. However, none of the existing interest point detectors is invariant to affine transformations. In this paper we present an affine invariant interest point detector. For each interest point we simultaneously adapt location as well as scale and shape of the neighbourhood. We then obtain a truly affine invariant image description which gives excellent results in the presence of arbitrary viewpoint changes. Note that a perspective transformation of a smooth surface can be locally approximated by an affine transformation.

1.1 Related work

Feature detection. Interest points are local features for which the signal changes two-dimensionally. They can be extracted reliably, are robust to partial visibility and the information content in these points is high. One of the first recognition techniques based on interest points has been proposed by Schmid and Mohr [14]. The points are extracted with the Harris detector [5] which is invariant to image rotation. To obtain invariance to scale changes interest points can be extracted in the scale space of an image [7]. Dufournaud et al. [3] use a multi-scale framework to match images at different scales. Interest points and descriptors are computed at several scales. A robust matching algorithm allows to select the correct scale. In the context of recognition, the complexity of a multi-scale approach is prohibitive. Lowe [10] proposes an efficient algorithm for recognition based on local extrema of difference-of-Gaussian filters in scale-space. Mikolajczyk and Schmid [11] use a multi-scale framework to detect points and then apply scale selection [8] to select characteristic points. These points are invariant to scale changes and allow matching and recognition in the presence of large scale factors. Tuytelaars and Van Gool [16] detect affine invariant regions based on image intensities. However, the number of such regions in an image is limited and depends on the content. They use colour descriptors computed for these regions for wide baseline matching.

Wide baseline matching and recognition. The methods presented in the following use standard feature detectors. They rely on the accuracy of these features which is a limitation in the presence of significant transformations.

Pritchett and Zisserman [12] estimate homographies of local planar surfaces in order to correct the cross-correlation and grow regions. The homographies are obtained by matching regions bound by four line segments. This approach has been applied to wide baseline matching and it is clearly difficult to extend to retrieval. Tell and Carlsson [15] also address the problem of wide baseline matching and use an affine invariant descriptors for point pairs. They compute an affine invariant Fourier description of the intensity profile along a line connecting two points. The description is not robust unless the two points lie on the same planar surface. Baumberg [2] extracts interest points at several scales and then adapts the shape of the region to the local image structure using an iterative procedure based on the second moment matrix [9]. Their descriptors are affine invariant for fixed scale and location, that is the scale and the location of the points are not extracted in an affine invariant way. The points as well as the associated regions are therefore not invariant in the presence of large affine transformations, see section 3.3 for a quantitative comparison to our approach. Furthermore, approximately four times more points are detected in comparison to our method. This increases the probability of false matches and in the case of retrieval the complexity is prohibitive. In our approach points that correspond to the same physical structure, but are detected at different locations in scale space, converge to the same point location. The number of points is therefore reduced. The properties of the second moment matrix were also explored by Schaffalitzky and Zisserman [13], but their goal was to obtain an affine invariant texture descriptor.

1.2 Our approach

A uniform Gaussian scale-space is often used to deal with scale changes [3, 7, 10, 11]. However, an affine Gaussian scale-space is too complex to be practically useful, as three parameters have to be determined simultaneously. In this paper we propose a realistic solution which limits the search space to the neighbourhood of points and uses an iterative search procedure. Our approach is based on a method introduced by Lindeberg and Garding [9] which iteratively estimates an affine invariant neighbourhood. They explore the properties of the second moment descriptor to recover the surface orientation and compute the descriptors with non uniform Gaussian kernels.

Our affine invariant interest point detector is an affine-adapted version of the Harris detector. The affine adaptation is based on the second moment matrix [9] and local extrema over scale of normalized derivatives [8]. Locations of interest points are detected by the affine-adapted Harris detector. For initialization, approximate localizations and scales of interest points are extracted by the multi-scale Harris detector. For each point we apply an iterative procedure which modifies position as well as scale and shape of the point neighbourhood. This allows to converge toward a stable point that is invariant to affine transformations. This detector is the main contribution of the paper. Furthermore, we have developed a repeatability criterion which takes into account the point position as well as the shape of the neighbourhood. A quantitative comparison with existing detectors [2, 11] shows a significant improvement of our method in the presence of large affine transformations. Results for wide baseline matching and recognition based on our affine invariant points are excellent in the presence of significant changes in viewing angle and scale and clearly demonstrate their invariance.

Overview. This paper is organized as follows. Section 2 introduces the key ideas of our approach. In section 3 our affine invariant interest point detector is described in detail and compared to existing approaches. The matching and recognition algorithm is outlined in section 4. Experimental results are given in section 5.

2 Affine Gaussian scale-space

In this section we extend the idea of searching interest points in the scale space representation of an image and propose to search points in an affine Gaussian scale space. We extend the approach proposed in [11]; this approach explores the properties of the uniform Gaussian scale space and can handle significant scale changes. It is based on interest points which are local maxima of the Harris measure above a threshold. The Harris measure is the second moment matrix and describes the gradient distribution in a local neighbourhood of a point $\mathbf{x}$:

$$\mu(\mathbf{x}, \sigma_I, \sigma_D) = \sigma_D^2 g(\sigma_I) * \begin{bmatrix} L_x^2(\mathbf{x}, \sigma_D) & L_x L_y(\mathbf{x}, \sigma_D) \\ L_x L_y(\mathbf{x}, \sigma_D) & L_y^2(\mathbf{x}, \sigma_D) \end{bmatrix} \tag{1}$$

$$\det(\mu) - \alpha \mathrm{trace}^2(\mu) > threshold \tag{2}$$

where σ_I is the integration scale, σ_D the derivation scale, g the Gaussian and L the image smoothed by a Gaussian (cf. equation 3). To deal with significant scale changes, points are extracted at several scales and the characteristic scale is determined by automatic scale selection [8]. Scale selection is based on the maximum of the normalized Laplacian $|\sigma^2(L_{xx}(\mathbf{x}, \sigma) + L_{yy}(\mathbf{x}, \sigma))|$ where derivatives are computed with uniform Gaussian filters. A problem occurs in the case of affine transformations where the scale changes are not necessarily the same in all directions. In this case the selected scale does not reflect the real transformation of a point. It is well known that the local Harris maxima have different spatial locations when extracted at different detection scales (see figure 1). Thus, an additional error is introduced to the location of the point if the detection scales do not correspond to the scale factor between corresponding image patterns. In the case of affine transformations the detection scales in x and y directions have to vary independently to deal with possible affine scaling. Suppose both scales can be adapted to the local image structure. Hence, we face the problem of computing the second moment matrix in affine Gaussian scale space, where a circular window is replaced by an ellipse. An affine scale-space can be generated by convolution with non-uniform Gaussian kernels:

$$g(\Sigma) = \frac{1}{2\pi\sqrt{det\Sigma}} \exp^{-\frac{\mathbf{x}^T \Sigma^{-1} \mathbf{x}}{2}},$$

where $\mathbf{x} \in \mathcal{R}^2$. If the matrix Σ is equal to an identity matrix multiplied by a scalar, this function corresponds to a uniform Gaussian kernel. Given any image function $I(\mathbf{x})$ the derivatives can be defined by

$$L_x(\mathbf{x}; \Sigma) = \frac{\partial}{\partial x} g(\Sigma) * I(\mathbf{x}) \tag{3}$$

This operation corresponds to the convolution with a rotated elliptical Gaussian kernel. If traditional uniform Gaussian filters are used, we deal with a three dimensional space (x, y, σ), and the Gaussian kernel is determined by one scale parameter σ. If Σ is a symmetric positive definite 2x2 matrix, the number of degrees of freedom of the kernel is three, which leads to a complex high dimensional search space. Thus, we have to apply additional constraints to reduce the search.

The selection of detection scales can be based on the second moment matrix. For a given point $\mathbf{x}$ the second moment matrix μ in non-uniform scale space is defined by

$$\mu(\mathbf{x}, \Sigma_I, \Sigma_D) = g(\Sigma_I) * ((\nabla L)(\mathbf{x}, \Sigma_D)(\nabla L)(\mathbf{x}, \Sigma_D)^T)$$

where Σ_I and Σ_D are the covariance matrices which determine the integration and the derivation Gaussian kernels. To reduce the search space we impose the

condition $\Sigma_I = a\Sigma_D$, where a is a scalar.
Consider an affine transformed point $\mathbf{x}_L = A\mathbf{x}_R$, the matrices μ are related by

$$\mu(\mathbf{x}_L, \Sigma_{I,L}, \Sigma_{D,L}) = A^T \mu(A\mathbf{x}_R, A\Sigma_{I,L}A^T, A\Sigma_{D,L}A^T)A \tag{4}$$

Lindeberg [9] showed that if the second moment descriptor of the point $\mathbf{x}_L$ verifies

$$\mu(\mathbf{x}_L, \Sigma_{I,L}, \Sigma_{D,L}) = M_L \qquad \Sigma_{I,L} = tM_L^{-1} \qquad \Sigma_{D,L} = dM_L^{-1}$$

and the descriptor of the point $\mathbf{x}_R$ verifies corresponding conditions

$$\mu(\mathbf{x}_R, \Sigma_{I,R}, \Sigma_{D,R}) = M_R \qquad \Sigma_{I,R} = tM_R^{-1} \qquad \Sigma_{D,R} = dM_R^{-1}$$

then the matrices M_L and M_R are related by

$$M_L = A^T M_R A \qquad A = M_R^{-1/2} R M_L^{1/2} \qquad \Sigma_R = A\Sigma_L A^T \tag{5}$$

where R is an arbitrary rotation. Note that the scalars t and d are the integration and derivation scales respectively. The relation 5 verifies equation 4. The proof and the outline of an iterative method for computing the matrices can be found in [9]. Matrices M_L and M_R, computed under these conditions, determine corresponding regions defined by $\mathbf{x}^T M \mathbf{x} = 1$. Baumberg [2] shows that if the neighbourhoods of points $\mathbf{x}_L$, $\mathbf{x}_R$ are normalized by transformations $\mathbf{x}'_L \mapsto M_L^{-1/2}\mathbf{x}_L$ and $\mathbf{x}'_R \mapsto M_R^{-1/2}\mathbf{x}_R$ respectively, then the normalized regions are related by a pure rotation $\mathbf{x}'_L \mapsto R\mathbf{x}'_R$. In the normalized frames M'_L and M'_R are equal up to a pure rotation matrix. In other words, the intensity patterns in the normalized frames are isotropic. We extend the approach proposed in [11]. We first transform the image locally to obtain an isotropic region and then search for a local Harris maximum and a characteristic scale. We then obtain a method for detecting points and regions invariant to affine transformations.

3 Affine invariant point detector

In order to limit the search space we initialize the affine detector with interest points extracted by the multi-scale Harris detector [3]. Any detector can be used to determine the *spatial localization* of the initial points. However, the Harris detector is based on the second moment matrix, and therefore naturally fits into our framework. To obtain the *shape adaptation matrix* for each interest point we compute the second moment descriptor with automatically selected *integration* and *derivation* scale. The outline of our detection method is presented in the following:

- the *spatial localization* of an interest point for a given scale and shape is determined by the affine-adapted Harris detector,
- the *integration scale* is selected at the extremum over scale of normalized derivatives,
- the *derivation scale* is selected at the maximum of normalized isotropy,
- the *shape adaptation matrix* normalizes the point neighbourhood.

In the following we discuss in detail each step of the algorithm.

Shape adaptation matrix. Our iterative shape adaptation method works in the transformed image domain. Instead of applying an adapted Gaussian kernel we can transform the image and apply a uniform kernel. A recursive implementation of the uniform Gaussian filters can then be used for computing L_x and L_y. The second moment matrix is computed according to equation 1. A local window is transformed by $U^{(k-1)} = (\mu^{-\frac{1}{2}})^{(k-1)} \cdots (\mu^{-\frac{1}{2}})^{(1)} \cdot U^{(0)}$ in step (k) of the iterative algorithm. In the following we refer to this operation as U-transformation. Note that a new μ matrix is computed at each iteration and that the U matrix is the concatenation of square roots of the second moment matrices. By keeping the larger eigenvalue $\lambda_{max}(U) = 1$ we assure that the original image is not under-sampled. This implies that the image patch is enlarged in the direction of $\lambda_{min}(U)$. For a given point the integration and the derivation scale determine the second moment matrix μ. These scale parameters are automatically detected in each iteration step. Thus, the resulting μ matrix is independent of the initial scale.

Integration scale. For a given spatial point we can automatically select its characteristic scale. In order to preserve invariance to scale changes we select the integration scale σ_I for which the normalized Laplacian $|\sigma^2(L_{xx}(\sigma) + L_{yy}(\sigma))|$ attains a local maximum over scale [9]. Keeping this scale constant during iterations can be sufficient in the presence of weak affine distortions. In the case of large affine deformations the scale change is in general very different for the x and y directions. Thus, the characteristic scale detected in the image domain and in its U-transformed version can be significantly different. It is, therefore, essential to select the integration scale after each estimation of the U transformation. This allows to converge towards a solution where the scale and the second moment matrix do not change any more.

Derivation scale. The local derivation scale is less critical and can be set proportional to the integration scale $\sigma_D = s\sigma_I$. The factor s should not be too small, otherwise the smoothing is too large with respect to the derivation. On the other hand s should be small enough such that σ_I can average the covariance matrix $\mu(\mathbf{x}, \sigma_D, \sigma_I)$ by smoothing. Factor s is commonly chosen from the range $[0.5, \ldots, 0.75]$. Our solution is to select the derivation scale for which the local isotropy assumes a maximum over this range of scales. The local isotropy is measured by the local gradient distribution μ (equation 1). To obtain a normalized measure we use the eigenvalue ratio $(\lambda_{min}(\mu)/\lambda_{max}(\mu))$. Given the integration scale σ_I we select $s \in [0.5, \ldots, 0.75]$ for which the ratio assumes a maximum. The factor s has an important influence on the convergence of the second moment matrix. The iterative procedure converges toward a matrix with equal eigenvalues. The smaller the difference between the eigenvalues $(\lambda_{max}(\mu), \lambda_{min}(\mu))$ of the initial matrix, the closer is the final solution and the faster the procedure converges. Note that the Harris measure (equation 2) already selects the points with two large eigenvalues. A large difference between the eigenvalues leads to a large scaling in one direction by the U-transformation and the point does not converge to a stable solution due to noise. Thus, the selection of the local scale allows to obtain a reasonable eigenvalue ratio and allows convergence for points which would not converge if the ratio is too large. A similar approach for local scale selection was proposed in [1].

Spatial localization. It is well known that the local maxima of the Harris measure (equation 2) change their spatial location if the detection scale changes. This can also be observed if the scale change is different in each direction. The detection with different scales in x and in y direction is replaced by affine normalizing the image and then applying the same scale in both directions. The affine normalization of a point neighbourhood slightly changes the local spatial maxima of the Harris measure. Consequently, we re-detect the maximum in the affine normalized window W. We then obtain a vector of displacement to the nearest maximum in the U-normalized image domain. The location of the initial point is corrected with the displacement vector back-transformed to the original image domain $\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} + U^{(k-1)} \cdot (\mathbf{x}_w^{(k)} - \mathbf{x}_w^{(k-1)})$, where $\mathbf{x}_w$ are the coordinates in the transformed image domain.

Termination criterion. The important part of the iteration procedure is the termination criterion. The convergence measure can be based on either the μ or the U matrix. If the criterion is based on the μ matrix computed in each iteration step, we require that this matrix is sufficiently close to a pure rotation matrix. This implies that $\lambda_{max}(\mu)$ and $\lambda_{min}(\mu)$ are equal. In practice we allow for a small error $\lambda_{min}(\mu)/\lambda_{max}(\mu) > \epsilon_C$. Another possibility is to interpret the transformation $U = R^T \cdot D \cdot R$ as a rotation R and a scaling D and compare consecutive transformations. We stop the iteration if the consecutive R and D transformations are sufficiently similar. Both termination criteria give the same final results. Another important point is to stop the procedure in the case of divergence. We reject the point if $\lambda_{max}(D)/\lambda_{min}(D) > \epsilon_l$ (i.e. $\epsilon_l = 6$), otherwise it leads to unstable elongated structures.

3.1 Detection algorithm

We propose an iterative procedure that allows initial points to converge to affine invariant points. To initialize our algorithm we use points extracted by the multi-scale Harris detector. These points are not affine invariant due to a non adapted Gaussian kernel, but provide an approximate localization and scale for initialization. For a given initial interest point $\mathbf{x}^{(0)}$ we apply the following procedure:

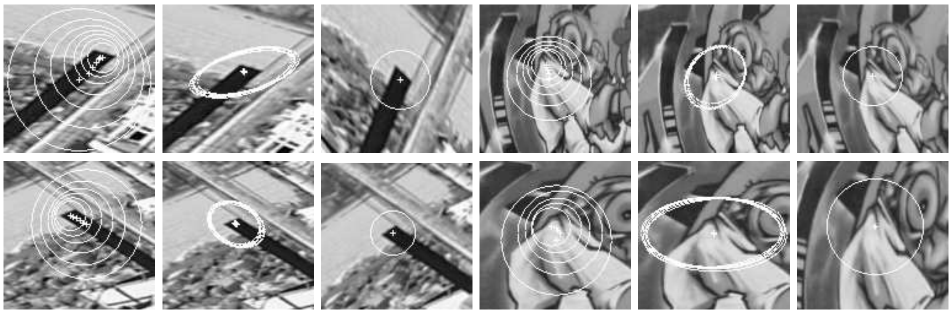
1. initialize $U^{(0)}$ to the identity matrix
2. normalize window $W(U^{(k-1)}\mathbf{x}_w) = I(\mathbf{x})$ centred in $U^{(k-1)}\mathbf{x}_w^{(k-1)} = \mathbf{x}^{(k-1)}$
3. select *integration scale* σ_I in $\mathbf{x}_w^{(k-1)}$
4. select *derivation scale* $\sigma_D = s\sigma_I$ which maximizes $\frac{\lambda_{min}(\mu)}{\lambda_{max}(\mu)}$ with $s \in [0.5, \ldots, 0.75]$ and $\mu = \mu(\mathbf{x}_w{}^{(k-1)}, \sigma_D, \sigma_I)$
5. detect *spatial localization* $\mathbf{x}_w^{(k)}$ of the maximum of the Harris measure (equation 2) nearest to $\mathbf{x}_w^{(k-1)}$ and compute the location of interest point $\mathbf{x}^{(k)}$
6. compute $\mu_i^{(k)} = \mu^{-\frac{1}{2}}(\mathbf{x}_w^{(k)}, \sigma_D, \sigma_I)$
7. concatenate transformation $U^{(k)} = \mu_i^{(k)} \cdot U^{(k-1)}$ and normalize $U^{(k)}$ such that $\lambda_{max}(U^{(k)}) = 1$
8. go to step 2 if $\lambda_{min}(\mu_i^{(k)})/\lambda_{max}(\mu_i^{(k)}) < \epsilon_C$

Although the computation may seem to be very time consuming, note that most time is spent computing L_x and L_y, which is done only once in each step if the factor s is kept constant. The iteration loop begins with selecting the integration scale because we have noticed that this part of the algorithm is most robust to a small localization error of an interest point. However, scale σ_I changes if the shape of the patch is transformed. Given an initial approximate solution, the presented algorithm allows to iteratively modify the shape, the scale and the spatial location of a point and converges to a true affine invariant interest point.

The convergence properties of the shape adaptation algorithm are extensively studied in [9]. In general the procedure converges provided that the initial estimation of the affine deformation is sufficiently close to the true deformation and that the integration scale is well adapted to the local signal structure.

3.2 Affine invariant interest point

Figure 1 presents two examples for interest point detection. Columns (a) display the points used for initialization which are detected by the multi-scale Harris detector. The circle around a point shows the scale of detection (the radius of the circle is $3\sigma_I$). Note that there is a significant change in location between points detected at different scales and that the circles in corresponding images (top and bottom row) do not cover the same image regions. The affine invariant points to which the initial points converge are presented in the columns (b). We can see that the method converges correctly even if the location and scale of the initial point is relatively far from the point of convergence. Convergence is in general obtained in less than 10 iterations. The minor differences between the



(a) (b) (c) (a) (b) (c)

Fig. 1. Affine invariant interest point detection : (a) Initial interest points detected with the multi-scale Harris detector. (b) Points and corresponding affine regions obtained after applying the iterative algorithm. (c) Point neighbourhoods normalized with the estimated matrices to remove stretch and skew.

regions in columns (b) are caused by the imprecision of the scale estimation and the error ϵ_C. The relation between two consecutive scales is 1.2 and ϵ_C is set to 0.96. It is easy to identify these regions by comparing their locations, scales and second moment matrices and to keep only one of them. We then obtain a set of points where each one represents a different image location and structure.

Column (c) shows the points normalized with the estimated matrices to remove stretch and skew. We can clearly see that the regions correspond between the two images (top and bottom row).

3.3 Repeatability of detectors

A comparative evaluation of different detectors is presented in the following. We compare our Harris-Affine method with two similar approaches [2, 11]. Mikolajczyk and Schmid [11] have developed a scale invariant interest point detector. Interest points are extracted at several scales with the Harris detector. Characteristic points are selected at the maxima over scale of the Laplacian function. We refer to this detector as Harris-Laplace. Baumberg [2] extracts Harris interest points at several scales and then adapts the shape of the region to the local image structure using an iterative procedure based on the second moment matrix. This method does not adapt location nor scale. It is referred to as Harris-AffineRegions.

An evaluation criterion for point detectors was described in [11]. It computes a repeatability score which takes into account the point location as well as the detection scale. We have extended this evaluation criterion to the affine case. The repeatability rate between two images is represented by the number of corresponding points with respect to the number of detected points. We consider two points $\mathbf{x}_a$ and $\mathbf{x}_b$ corresponding if :

1. the error in relative location of $\|\mathbf{x}_a, H \cdot \mathbf{x}_b\| < 1.5$ pixel, where H is the homography between images (planar scenes are used for our evaluation)
2. the error in image surface covered by point neighbourhoods is less than 20%

$$\epsilon_S = 1 - \frac{\mu_a \cap (A^T \mu_b A)}{\mu_a \cup (A^T \mu_b A)} < 0.2 \tag{6}$$

where μ_a and μ_b are the elliptical regions defined by $x^T \mu x = 1$. The union of the regions is $(\mu_a \cup (A^T \mu_b A))$ and $(\mu_a \cap (A^T \mu_b A))$ is their intersection. A is a local linearization of the homography H in point $\mathbf{x}_b$. We neglect the possible 1.5 pixel translation error while computing ϵ_S, because it has a small influence and the homography between real images is not perfect.

0° 40° 50° 60° 70°

Fig. 2. Images of one test sequence. The corresponding viewpoint angles are indicated below the images.

Figures 3 and 4 display average results for three real sequences of planar scenes (see figure 2). The viewpoint varied in horizontal direction between 0 and

70 degree. There are also illumination and zoom changes between the images. The homography between images was estimated with manually selected point pairs. Figure 3 displays the repeatability rate and figure 4 shows the localization

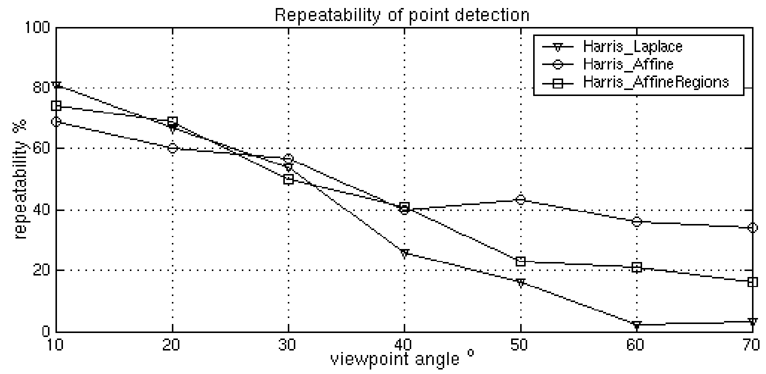


Fig. 3. Repeatability of detectors: a)$Harris\_Affine$ - approach proposed in this paper, b)$Harris\_AffineRegions$ - the multi-scale Harris detector with affine normalization of the point regions, c)$Harris\_Laplace$ - the multi-scale Harris detector with characteristic scale selection.

and intersection error for corresponding points. We can notice in figure 3 that our detector significantly improves the results for strong affine deformations, that is for changes in the viewpoint angle of more than 40 degrees. The improvement is with respect to localization as well as region intersection (see figure 4). In the presence of weak affine distortions the $Harris\_Laplace$ approach provides slightly better results. The affine adaptation does not improve the location and the point shape because the scaling is almost the same in every direction. In this case the uniform Gaussian kernel is sufficiently well adapted.

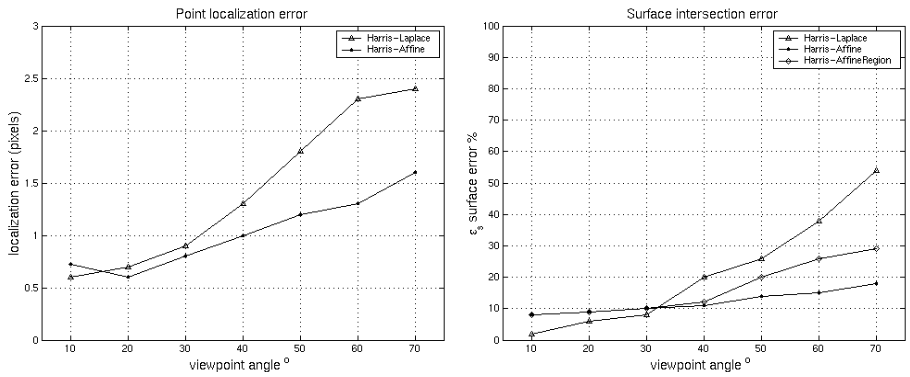


Fig. 4. Detection error of corresponding points : a) relative location b) surface intersection ϵ_S.

4 Matching and recognition

Point detection. The initial set of interest points is obtained with the multi-scale Harris detector. The scale-space representation starts with a detection scale of 2.5 and the scale factor between two levels of resolution is 1.2. We have used 17 scale levels. The parameter α is set to 0.06 and the threshold for the Harris detector is set to 1000 (cf. equation 2). For every point we then applied the iterative procedure to obtain affine invariant points. The allowed convergence error ϵ_C is set to 0.96. Similar points are eliminated by comparing location, scale and second moment matrices. About 40% of the points do not converge and 2/3 of the remaining points are eliminated by the similarity measure, that is 20-30% of initial points provided by the multi-scale Harris detector are kept.

Descriptors. Our descriptors are normalized Gaussian derivatives. Derivatives are computed on image patches normalized with the matrix U estimated for each point. Invariance to rotation is obtained by "steering" the derivatives in the direction of the gradient [4]. To obtain a stable estimation of the gradient direction, we use an average gradient orientation in a point neighbourhood. Invariance to affine intensity changes is obtained by dividing the derivatives by the first derivative. We obtain descriptors of dimension 12 by using derivatives up to 4th order.

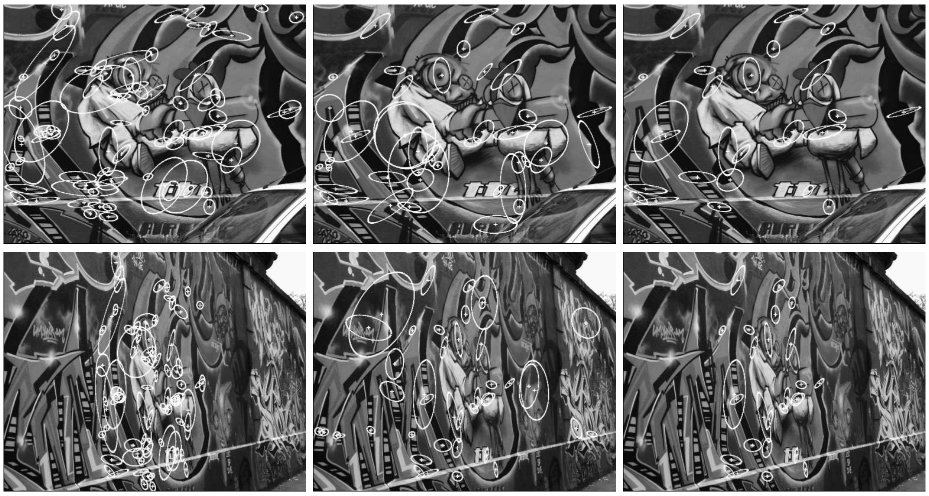
Similarity of descriptors. The similarity of descriptors is measured by the Mahalanobis distance. This distance requires the estimation of the covariance matrix Λ which encapsulates signal noise, variations in photometry as well as inaccuracy of the interest point location. Λ is estimated statistically over a large set of image samples. Given the scale, the gradient direction and the neighbourhood shape of points we can affine normalize the window and use cross-correlation as an additional distance measure.

Robust matching. To robustly match two images, we first determine point-to-point correspondences. We select for each descriptor in the first image the most similar descriptor in the second image based on the Mahalanobis distance. If the distance is below a threshold, the match is kept. We obtain a set of initial matches. These matches are verified by the cross-correlation measure which rejects less significant matches. Finally a robust estimation of the geometric transformation between the two images based on RANdom SAmple Consensus (RANSAC) rejects inconsistent matches. For our experimental results the transformation used is either a homography or a fundamental matrix. A model selection algorithm [6] can be used to automatically decide which transformation is the most appropriate one.

Database retrieval. A voting algorithm is used to select the most similar images in the database. This makes retrieval robust to mismatches as well as outliers. For each interest point of a query image, its descriptor is compared to the descriptors in the database. If the distance is less than a fixed threshold, a vote is added for the corresponding database image. Note that a point cannot vote several times for the same database image. The database image with the highest number of votes is the most similar one.

5 Experimental results

In this section we present matching and recognition results based on the method described in section 4. All the tests were carried on real images [17].



(a) (b) (c)

Fig. 5. Robust matching : (a) There are 78 couples of possible matches among the 287 and 325 detected points. (b) There are 43 point matches based on the descriptors and the cross-correlation score. 27 of these matches are correct. (c)There are 27 inliers to the robustly estimated homography. All of them are correct.

Matching. Figure 5 illustrates the results of the matching procedure. In order to separate the detection and matching results, we present all the possible correspondences determined with the homography in column (a). There are 78 corresponding point pairs among the 287 and 325 points detected in the first and second images respectively. We first match the detected points with the Mahalanobis distance and obtain 53 matches (29 correct and 24 incorrect). An additional verification based on the cross-correlation score rejects 10 matches (2 correct and 8 incorrect). These 43 matches (27 correct and 16 incorrect) are displayed in column (b). The images in column (c) show the 27 inliers to the robustly estimated homography. Note that there is a significant perspective transformation between the two images. A second example is presented in figure 6a. The images show a 3D scene taken from significantly different viewpoints. This image pair presents a more significant change in viewpoint than the images in figure 7c which were used in [13, 16] as an example for matching. In the figure 6b, we show a pair of images for which our matching procedure fails. The failure is not due to our detector, as the manually selected corresponding points show. It is caused by our descriptors which are not sufficiently distinctive. Note that the corners of sharp or wide angles, of light or dark intensity are almost the same once normalized to be affine invariant. If there is no distinctive texture in the region around the points, there are too many mismatches and additional constraints as for example semi-local constraints [3] should be used.

(a) b)

Fig. 6. (a) Example of a 3D scene observed from significantly different viewpoints. There are 14 inliers to a robustly estimated fundamental matrix, all of them correct. (b) An image pairs for which our method fails. There exist, however, corresponding points which we have selected manually.

Database retrieval. In the following we present retrieval results from a database with more than 5000 images. The images in the database are extracted from video sequences which include movies, sport events and news reports. Similar images are mostly excluded by taking one image per 300 frames. Furthermore, the database contains one image of each of our 4 test sequences. The second row of figure 7 shows these four images. The top row displays images for which the corresponding image in the database (second row) was correctly retrieved. Note the significant transformations between the query images and the images in the database. There is a scale change of a factor of 3 between images of pair (a). Image pairs (b) and (c) show large changes in viewing angle. Image pair (d) combines a scale change with a significant change in viewing angle. The displayed matches are the inliers to a robustly estimated transformation matrix between the query image and the most similar image in the database.

Conclusions and discussion

In this paper we have described a novel approach for interest point detection which is invariant to affine transformations. Our algorithm simultaneously adapts location as well as scale and shape of the point neighbourhood. None of the existing methods for interest point detection simultaneously solves for all of these problems during feature extraction. Our affine invariant points and the associated corresponding regions allow matching and recognition in the presence of large scale and viewpoint changes. Experimental results for wide baseline matching and recognition are excellent. Future work includes the development of more discriminant descriptors as well as the use of neighbourhood constraints.

(a) (b) (c)

(d)

Fig. 7. For image pairs (a),(b) and (c) the top row shows the query images and the bottom row shows the most similar images in the database. For image pair (d) the left image is the query image and the right one the image in the database. The displayed matches are the inliers to a robustly estimated fundamental matrix or homography between the query image and the most similar image in the database. There are (a) 22 matches, (b) 34 matches, (c) 22 matches and (d) 33 matches. All of them are correct.

Acknowledgement

This work was supported by the European FET-open project VIBES. We are grateful to RobotVis INRIA Sophia-Antipolis for providing the Valbonne images and to the University of Oxford for the Dunster images. The authors would like to express special thanks to David Lowe and Matthew Brown for useful suggestions and constructive discussions during a preliminary part of this work.

References

1. A. Almansa and T. Lindeberg. Fingerprint enhancement by shape adaptation of scale-space operators with automatic scale selection. IEEE *Transactions on Image Processing*, 9(12):2027–2042, 2000.
2. A. Baumberg. Reliable feature matching across widely separated views. In *Proceedings of the Conference on Computer Vision and Pattern Recognition, Hilton Head Island, South Carolina, USA*, pages 774–781, 2000.
3. Y. Dufournaud, C. Schmid, and R. Horaud. Matching images with different resolutions. In *Proceedings of the Conference on Computer Vision and Pattern Recognition, Hilton Head Island, South Carolina, USA*, pages 612–618, 2000.
4. W. Freeman and E. Adelson. The design and use of steerable filters. IEEE *Transactions on Pattern Analysis and Machine Intelligence*, 13(9):891–906, 1991.
5. C. Harris and M. Stephens. A combined corner and edge detector. In *Alvey Vision Conference*, pages 147–151, 1988.
6. K. Kanatani. Geometric information criterion for model selection. *International Journal of Computer Vision*, 26(3):171–189, 1998.
7. T. Lindeberg. *Scale-Space Theory in Computer Vision*. Kluwer Publishers, 1994.
8. T. Lindeberg. Feature detection with automatic scale selection. *International Journal of Computer Vision*, 30(2):79–116, 1998.
9. T. Lindeberg and J. Garding. Shape-adapted smoothing in estimation of 3-D shape cues from affine deformations of local 2-D brightness structure. *Image and Vision Computing*, 15(6):415–434, 1997.
10. D. G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the 7th International Conference on Computer Vision, Kerkyra, Greece*, pages 1150–1157, 1999.
11. K. Mikolajczyk and C. Schmid. Indexing based on scale invariant interest points. In *Proceedings of the 8th International Conference on Computer Vision, Vancouver, Canada*, pages 525–531, 2001.
12. P. Pritchett and A. Zisserman. Wide baseline stereo matching. In *Proceedings of the 6th International Conference on Computer Vision, Bombay, India*, pages 754–760, 1998.
13. F. Schaffalitzky and A. Zisserman. Viewpoint invariant texture matching and wide baseline stereo. In *Proceedings of the 8th International Conference on Computer Vision, Vancouver, Canada*, pages 636–643, 2001.
14. C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. IEEE *Transactions on Pattern Analysis and Machine Intelligence*, 19(5):530–534, 1997.
15. D. Tell and S. Carlsson. Wide baseline point matching using affine invariants computed from intensity profiles. In *Proceedings of the 6th European Conference on Computer Vision, Dublin, Ireland*, pages 814–828, 2000.
16. T. Tuytelaars and L. Van Gool. Wide baseline stereo matching based on local, affinely invariant regions. In *The Eleventh British Machine Vision Conference, University of Bristol, UK*, pages 412–425, 2000.
17. Test sequences. http://www.inrialpes.fr/movi/people/Mikolajczyk/Database/